DATA+AI
SUMMIT 2022

Building an Operational ML Team from Zero

& Leveraging Machine Learning for Crypto Security

ORGANIZED BY databricks

Anthony Tellez
Sr. Security Architect, BlockFi

Anthony Tellez

CISSP,CEH,CNDA

Head of Machine Learning at BlockFi

Previous:

- +6y Principal ML Architect, Splunk
- Geospatial Analyst, U.S. Gov

Specializations

- ML for Network Security (DGA & Bot Detection)
- Data Visualization for security and fraud analytics
- Applied Data Science, Graph Theory

Pre-pandemic: Avg ~150,000 Airline miles per year traveling around the world.

Agenda

Do more with your data!

Data & ML at BlockFi

BlockFi's data-centric culture is focused on using Machine Learning to provide world-class service to our clients and internal stakeholders.

Topics

- BlockFi Overview
- Building a Machine Learning Organization
- The Security Landscape for Cryptocurrency
- What is Graph Analytics?
- Threat Intelligence Mining using ML
- Using Graphs for Cryptotracing

Don’t just buy Bitcoin, earn it.

BlockFi

*terms apply

BlockFi's Vision & Strategy

BlockFi

We envision

a future where crypto-powered financial accelerate prosperity worldwide.

We are pioneering

the transformation of financial services, using blockchain rails to radically improve traditional financial products.

We are on a mission

to be the most trusted provider of crypto-powered financial services by bringing innovation, transparency, and compliance to digital asset markets worldwide.

BlockFi Retail

Do more with your crypto

BlockFi provides financial products for crypto investors. Current products include crypto accounts, USD loans backed by crypto, low cost trading, and the world's first crypto rewards credit card.

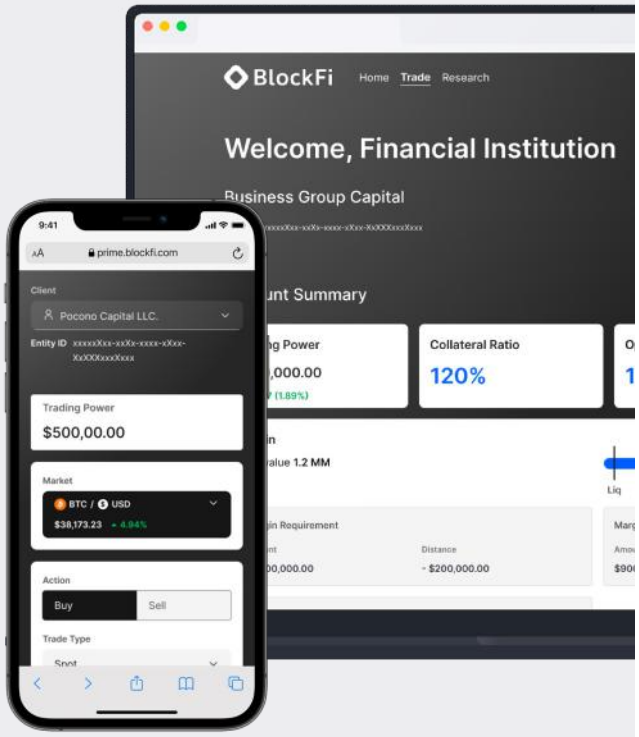
BlockFi Wallet	**Crypto - backed Lending**	**Trading**	**Credit Card** *US Clients only	**Personalized Yield**
Safely buy, sell and store crypto and stablecoins	Borrow USD using BTC, ETH, or LTC as collateral.	Trade BTC, ETH, LTC, or Stablecoin with low costs.	Visa Signature credit card, Earn crypto on every purchase.	Negotiate crypto interest rates, fiat borrowing, and trading costs to high net worth clients

BlockFi Institutional

Do more with your crypto

BlockFi is building a bridge between traditional finance and crypto - offering financial products and services to institutional clients, hedge funds, market makers, family offices and crypto native firms.

BlockFi Prime	Corporate Lending	Miners, exchanges & ATMs	Private client offering
Borrow digital assets or USD at negotiated terms and rates for hedging, market-making, shorting or for other working capital needs	Borrow USD or crypto to fund your business	Customize transaction terms to meet unique financing and hedging needs for mining equipment backed and crypto loans	Personalized crypto products and services along with 1:1 relationship manager for HNW and UHNW investors



Building an Operational Machine Learning Organization

Team Building 🚀

& Organizational Collaboration

Identify Stakeholders

Strong data science and engineering talent in Marketing, Fraud, Finance, but lacked a leader focused on data strategy and a secure platform for analytics.

Scope the Problem

Teams using the same processes to build brittle reporting capabilities. Collaboration capabilities between departments, de-duplication of data, and engineering work were impossible to identify. Security concerns and platform stability.

Collaborate, Build, Win.

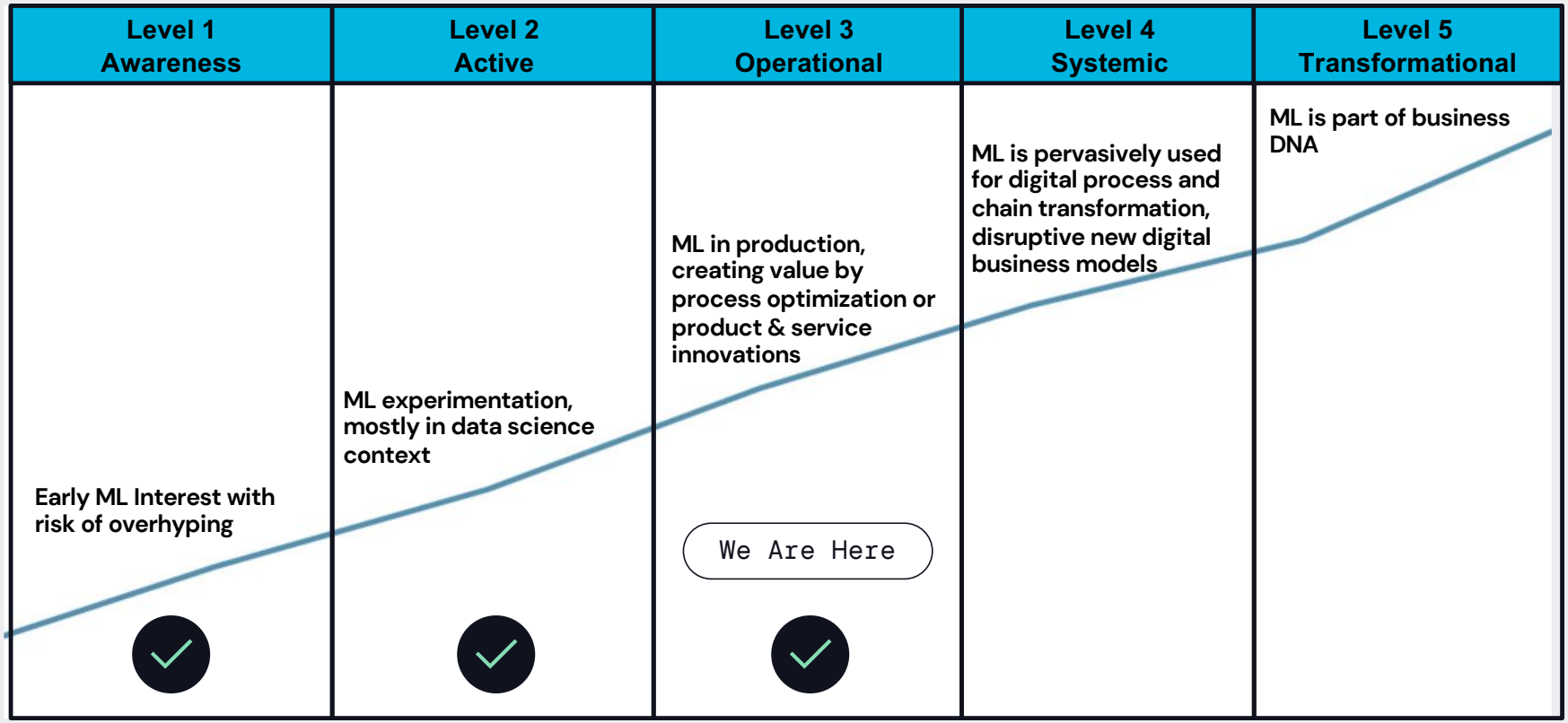
Identified easy wins with partners in Finance, Fraud & Marketing teams.

Ex: Analytic views that are required by business processes across the stakeholders for crypto transactions, leverage automation to eliminate manual data ETL processes.

Regularly communicate!

ML Maturity Model

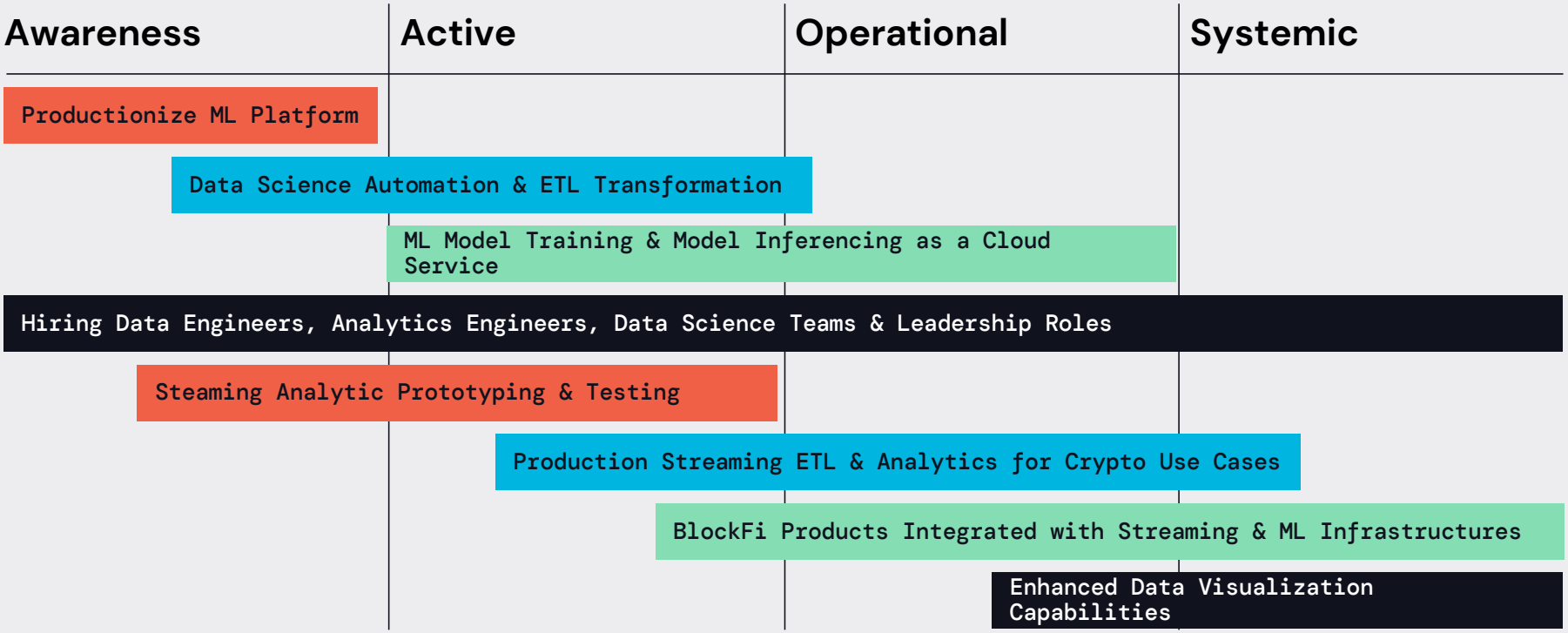
BlockFi's Journey to Operational ML & Beyond



Adapted from: Gartner, Five Stage AI Maturity Model: https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence

ML Roadmap

BlockFi's Journey to Operational ML & Beyond



Solving Security Challenges and Business Problems

Unique Problems in Crypto

`Crypto never sleeps.`

- Early in Regulatory Frameworks
- Speed of Activity
- 24/7 Live Trading
- Publically Connected Node Infrastructure
- 51% Attacks
- Methodology for Detecting Fraud & Dust Attacks

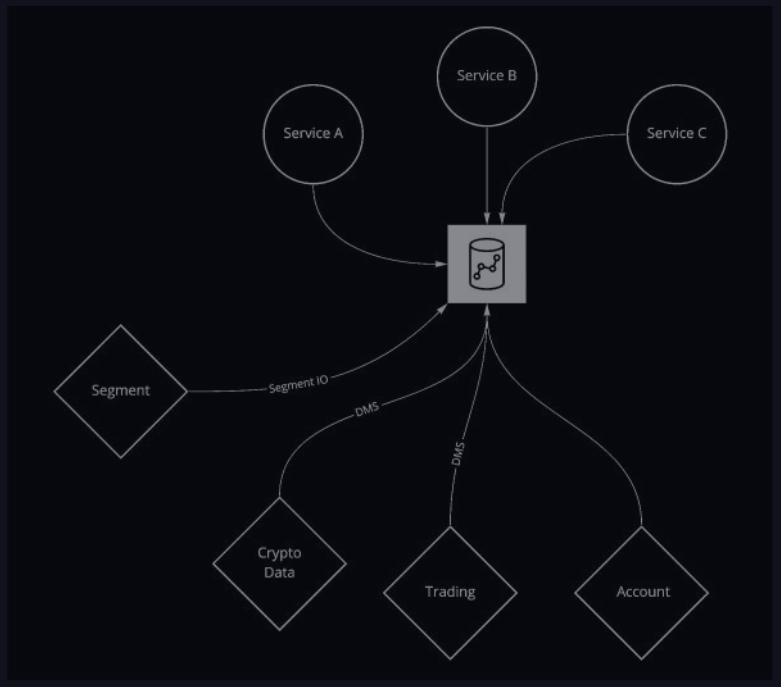
Historical Data Architecture

Redshift Warehouse Access Patterns and Challenges

DIY Data Architecture

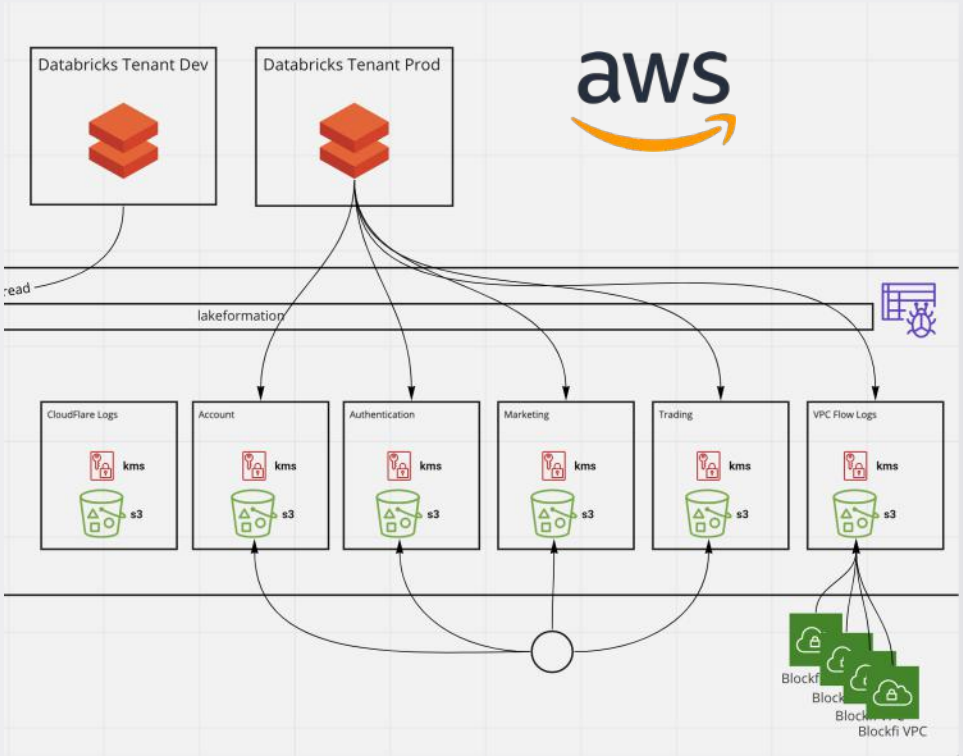
AWS Redshift, no clustering

- Scale out patterns relied on DMS replication
- User provisioning on a per resource basis using IAMs and static credentials
- Difficult to scope **Union** of access across tables, views and IAM roles
- **Data leakage** as end users would cache copies of data locally
- Lack of ***Infrastructure As Code*** for deployment



AWS LakeFormation

Optimizing AWS resources



S3 is cheap, scalable and secure.

AWS Lakeformation allows you to easily write parquet, csv and log files into s3 buckets with role based policies.

AWS native data is directly written into S3 for processing by Databricks

Infinitely scalable for storage and custom KMS keys are supported

3rd Party Vendors can be granted Read/Write ACLs to upload data into the Lake using Cross-Account trust

LakeHouse Architecture

`Okta + DeltaLake =` 🔥🔥🔥

Data Refinery

Data landing in AWS LakeFormation is considered **Raw** and needs refinement before it's ready for analytics and machine learning.

Leverage DeltaLake ETL to build new features, enrich and validate data.

Raw > Bronze > Silver >

Feature Orchestration

ETL jobs leverage native job scheduler and Airflow to orchestrate more complex joins of data across data stores to build data for reporting and machine learning.

Security Made Easier

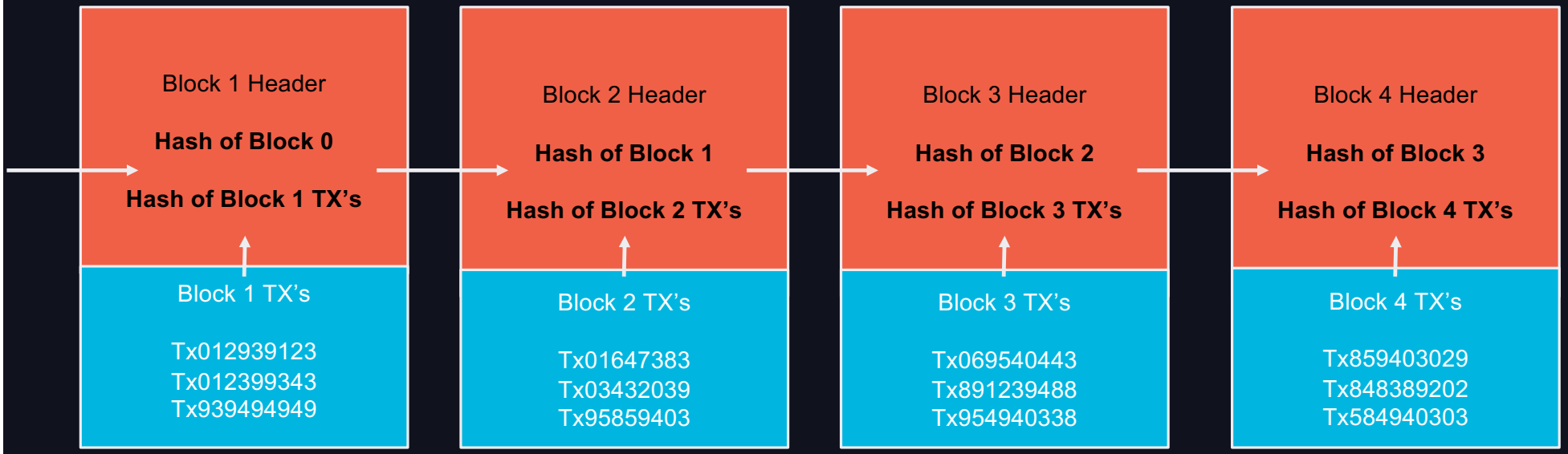
Group level ACLs leveraged by importing membership from SSO & SCIM.

Users have group based access to Data, Compute and Business Critical reports. Easy to validate, audit, and revoke.

Security Landscape for Cryptocurrency

Blockchain TX

Example blockchain flow

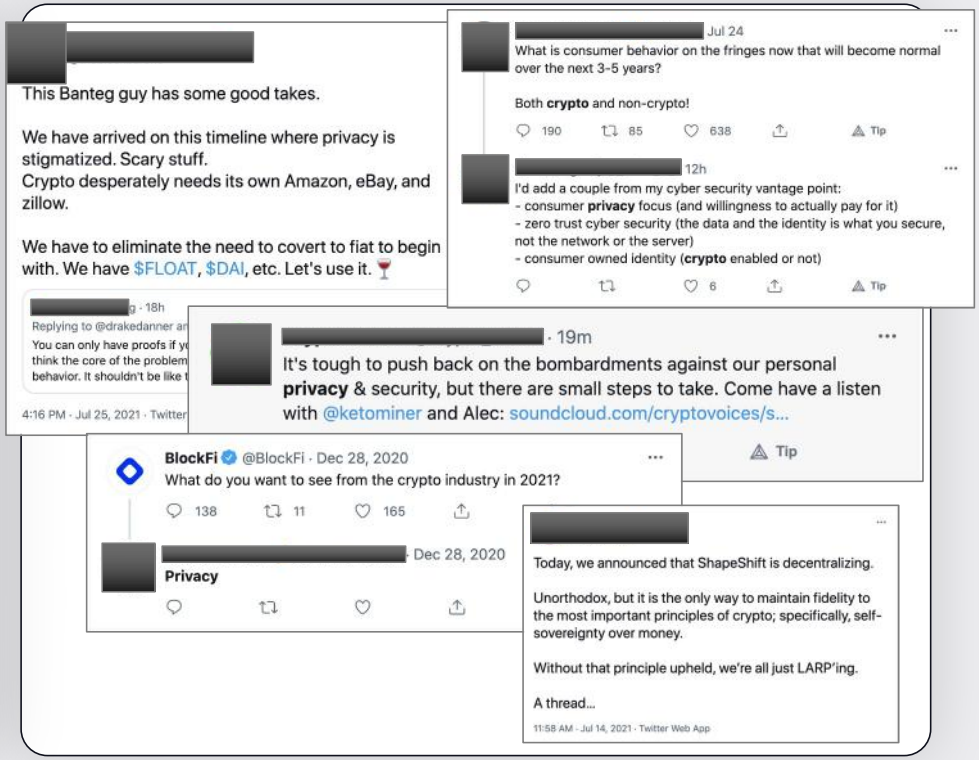


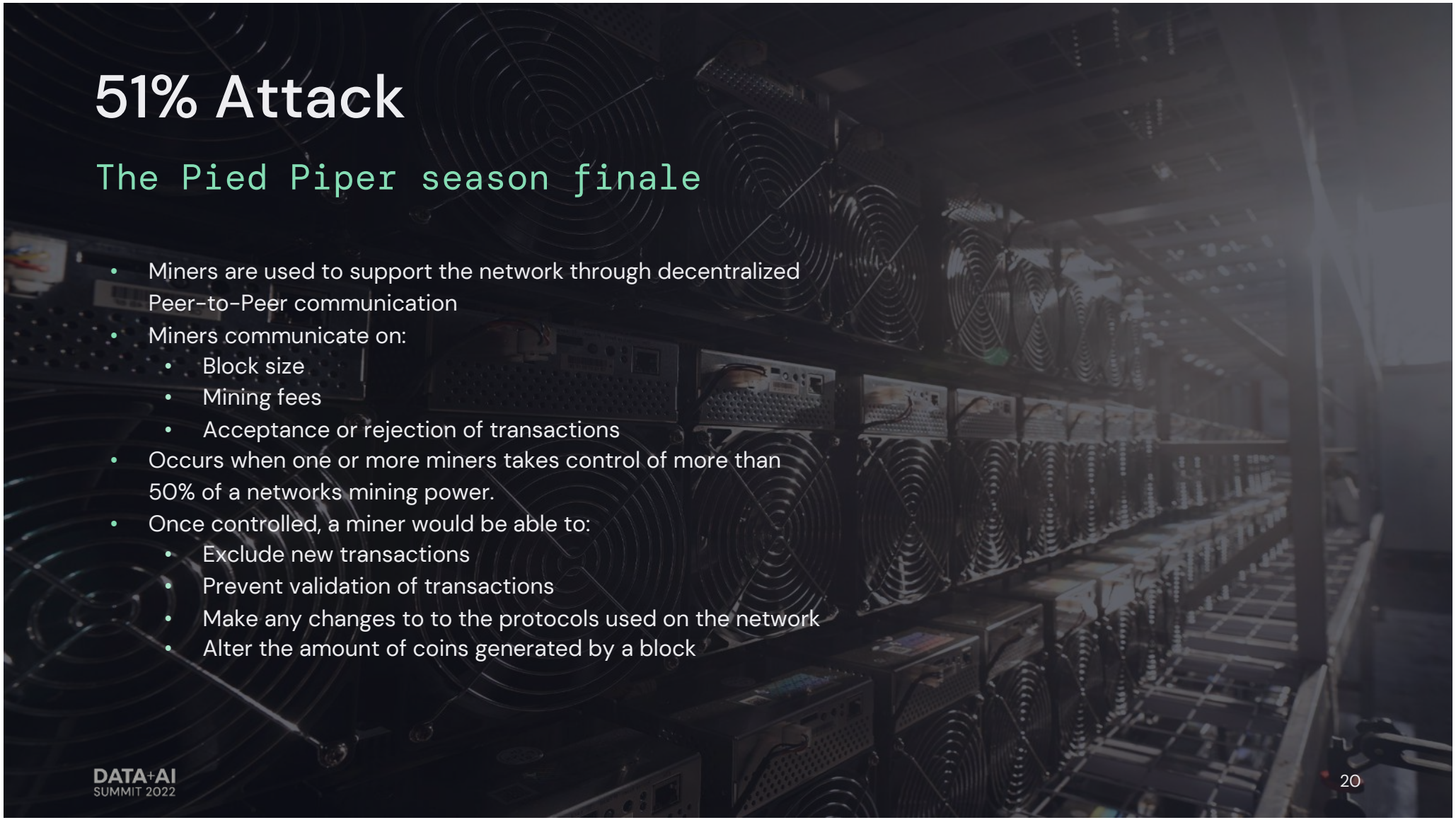
Crypto-Exchange Attack Surface

Cutting through the FUD & Hype

Industry at a Glance

- Young Industry
  - Largely values privacy
  - Technologically Savvy
  - Immature products and business processes
- Quick money movement
  - No take backs, once the transaction is confirmed
- Early in Regulatory Frameworks
- Largely immature in security-posture and response
- Exchange Requires at minimum:
  - Crypto node Infrastructure
  - Private-Key Infrastructure
- Targeted by attack groups for fraud and to rinse ransomware bounty
- Big presence on social media





51% Attack

The Pied Piper season finale

- Miners are used to support the network through decentralized Peer-to-Peer communication
- Miners communicate on:
  - Block size
  - Mining fees
  - Acceptance or rejection of transactions
- Occurs when one or more miners takes control of more than 50% of a networks mining power.
- Once controlled, a miner would be able to:
  - Exclude new transactions
  - Prevent validation of transactions
  - Make any changes to to the protocols used on the network
  - Alter the amount of coins generated by a block

Dusting Attacks

Type of reconnaissance used to de-anonymize or track potential wallet addresses.

Plan

Actor takes an interest in finding targets of opportunity or surveillance

Attack

Small amounts of crypto sent to randomly generated addresses using script/bot

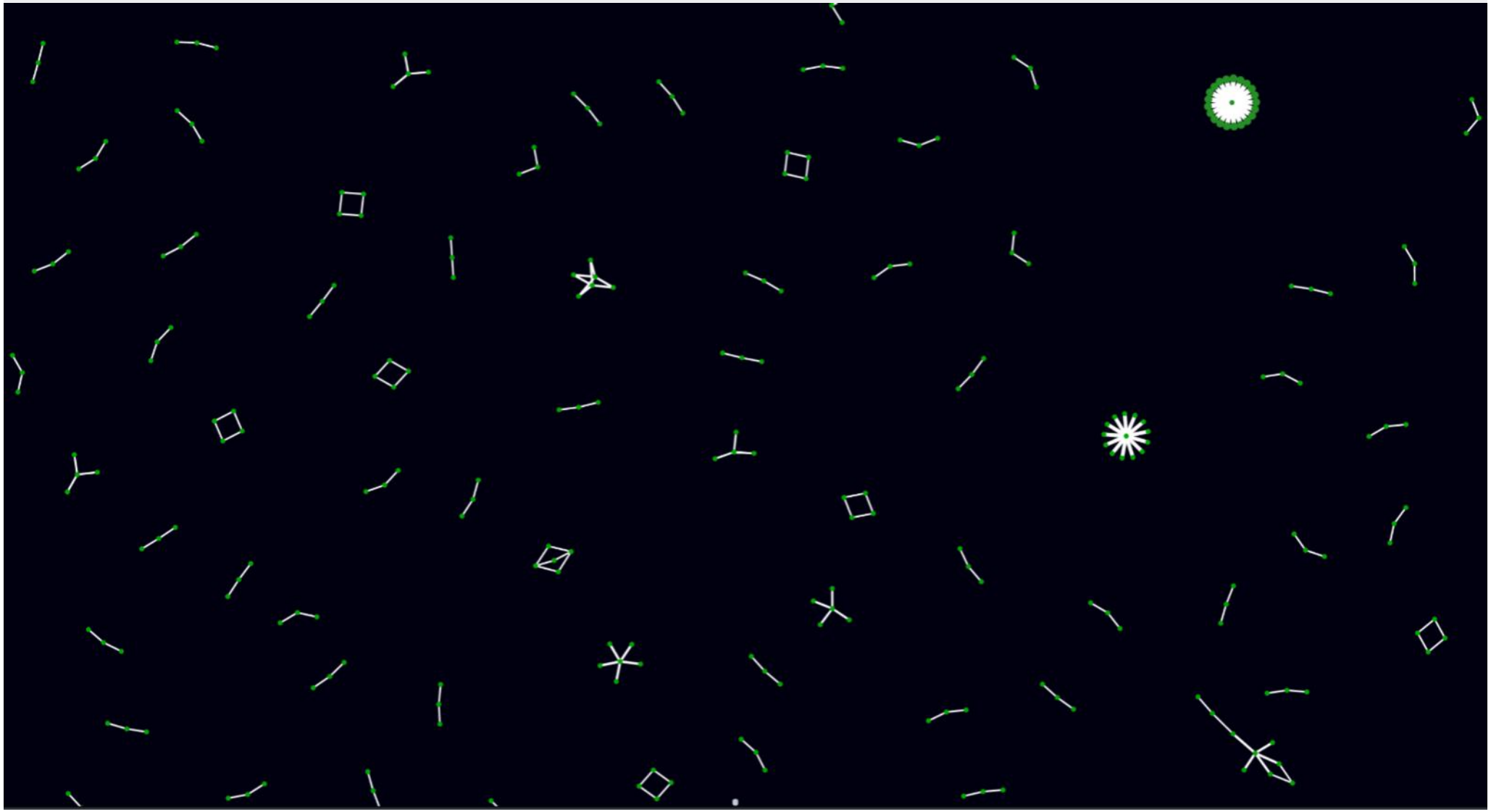
Refine

Valid wallets are kept as list and wallets with large balances filtered

Monitor/Exploit

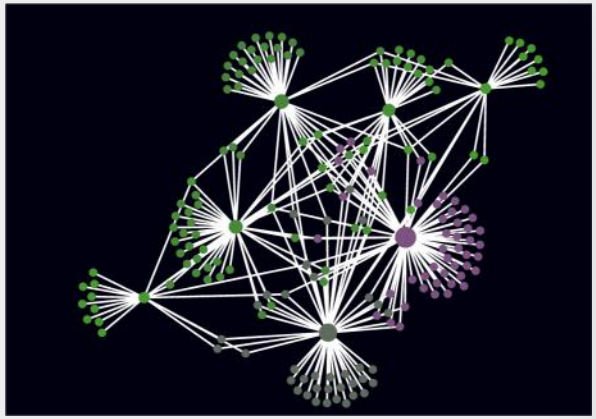
Public ledger used to track the movement of funds from wallets or attempt scams.

Graph Analytics and Six Degrees of Separation



Building Graphs

What is a graph?



"In mathematics, graph theory is the study of graphs, which are mathematical structures used to model pairwise relations between objects."

Why use a graph?

Graphs are generally leveraged to identify important nodes within a network, but they can also be used to build recommendation engines in social media and e-commerce applications.

In security, graphs can be used to identify users or devices within a network with a high level of connectivity and identify shared properties of those nodes.

- Accounts with shared IP Addresses or Device
- Accounts with shared Wallet Addresses
- Accounts with shared behavioral patterns (session behavior)

Building a Graph Service

Example code

Example DataFrame

src	dest	amount	asset
0x123	0x987	0.03	btc
0xABC	0xZXY	10	eth
0x567	0x987	5000	doge

```
import graphistry
g = graphistry.edges(df, 'src'
'dest')
url = g.plot(render=False)
displayHTML(url)
```

Scalable + GPU Acceleration FTW

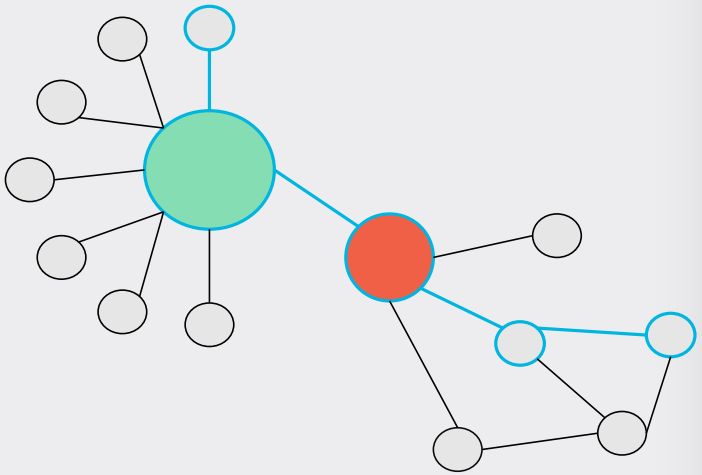
Graphistry leverages Nvidia Rapids (cuDF, cuGraph) to process data and build graphs.

This is beneficial when you have a hypergraph (eg: lots of connections) and cannot rely on NetworkX to process the graph in a reasonable amount of time.

Using Databricks we pass data to the API service and quickly build out the nodes and edges using a GPU. The result can be displayed in-line as a visualization in Databricks notebooks!

NetworkX & cuGraph Features

Simple Graph Example



Centrality Measurements

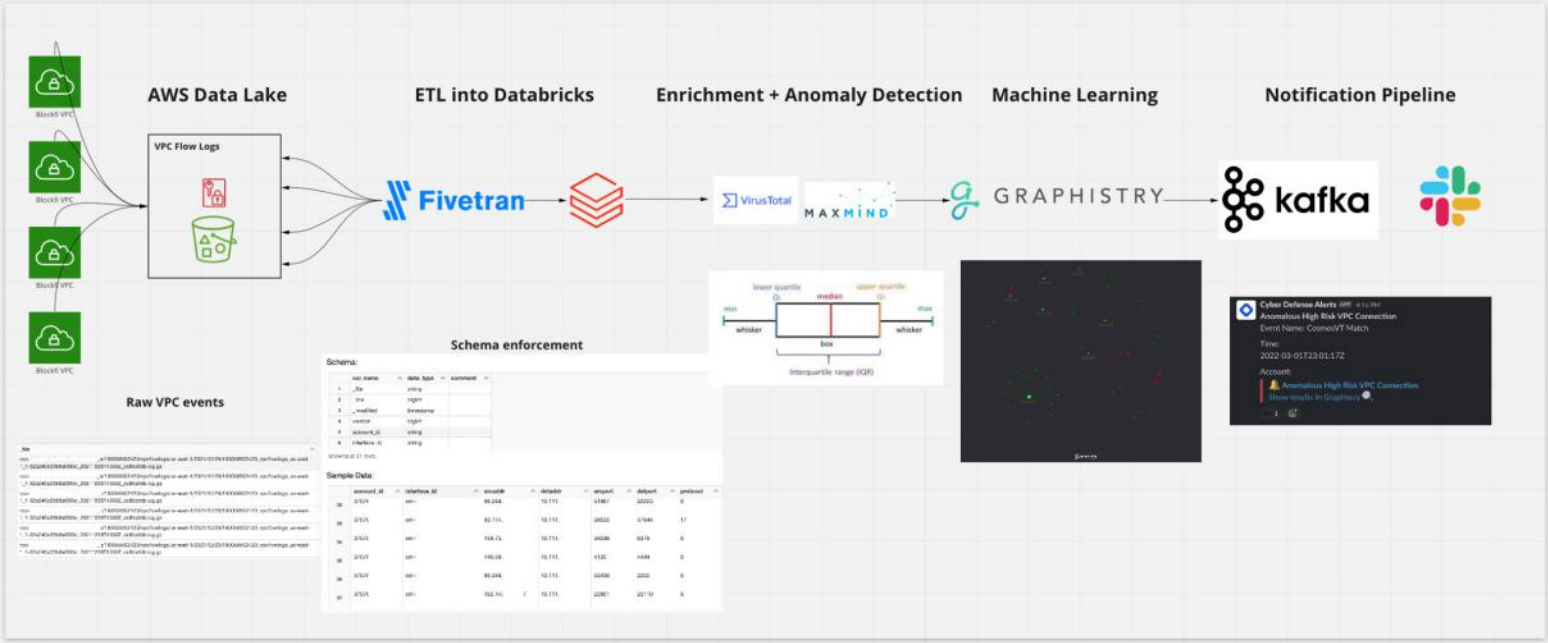
Can be used to identify different measures of a node's importance to the overall graph structure.

- **Eigenvector Centrality** - a measure of the influence of a node in a network.
- **Clustering Coefficient** - a measure of the degree to which nodes in a graph tend to cluster together.
- **Betweenness Centrality** - a measure of centrality in a graph based on shortest paths.

Threat Intelligence Mining using ML

Threat Intel Pipeline

With a side of anomaly detection & machine learning

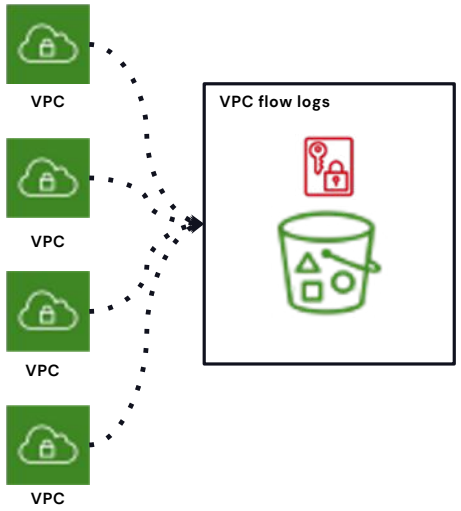


- Streamline 750GB of VPC log data per day to actionable insights for cyber security analysts to investigate

VPC logs

Understanding our netflows

AWS Data Lake



Raw VPC events

Know thy network

The collection of VPC flow logs is cumbersome but vital for BlockFi to diagnose, monitor, and understand our IP traffic going to and from our network interfaces

Diagnose + Monitor + Understand

Manage and validate security group rules to restrict open access

Ensure network traffic to our endpoints are what we originally allow them to be

Identify outliers and provide context to suspicious network behavior

ETL tools

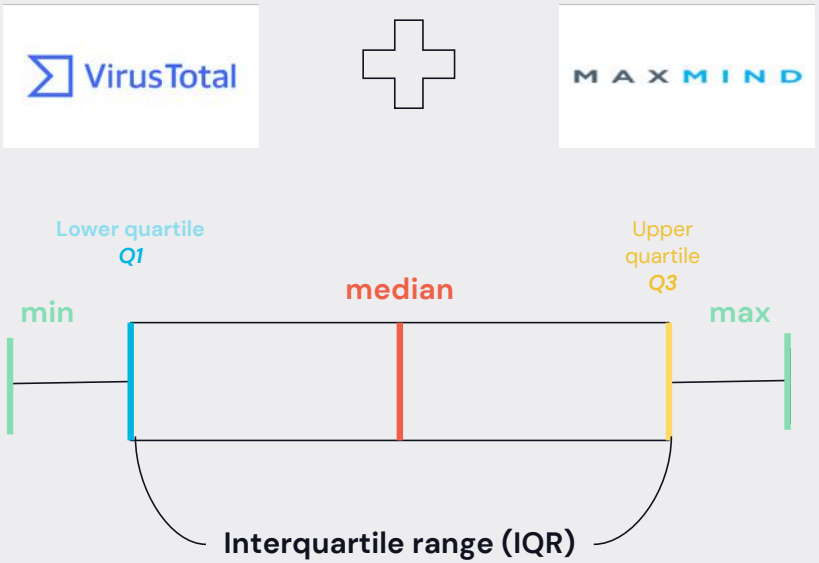
Pipeline beginnings

The initial aggregators of our pipeline must be robust to ensure data transfers are organized, reliable and setup for downstream enrichment

Organized	Reliability	Context
Manage and validate security group rules to restrict open access	Ensure network traffic to our endpoints are what we originally allow them to be	Identify outliers and provide context to suspicious network behavior

Enrichment + Anomaly Detection

Affirmation, Math and Labels



Context is king! Our security analysts must be empowered to make high pressure decisions as quickly as possible

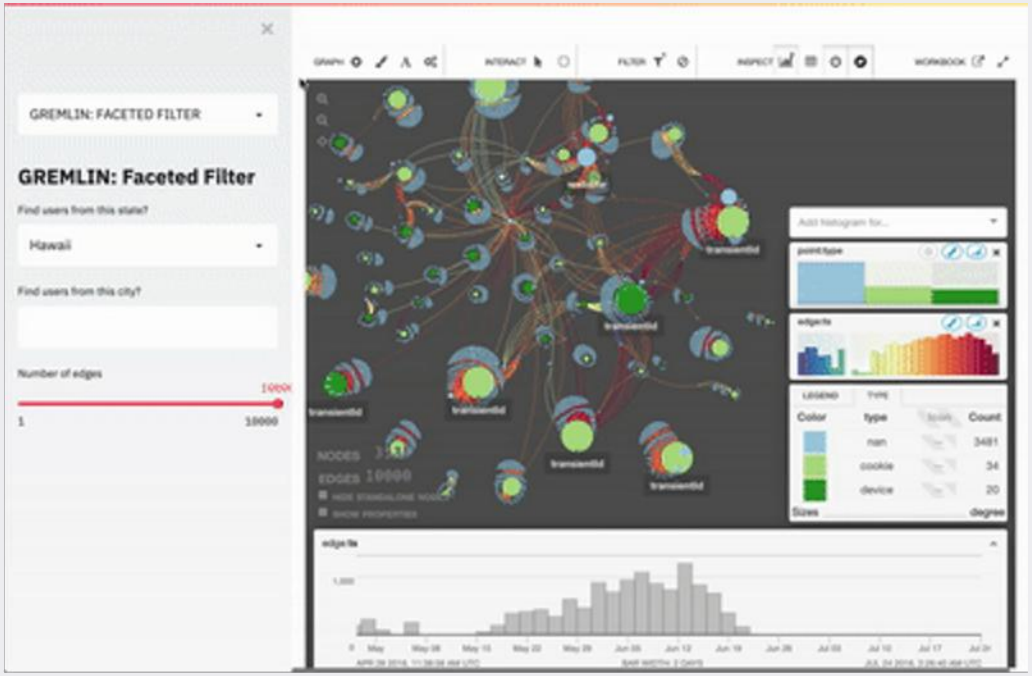
- **VirusTotal**
  - Rich details and affirmation on indicators of compromise
- **MaxMind**
  - Context on risky IP addresses connecting to our network
- **IQR**
  - Outlier detection on network features like # of packets, bytes, time of network communication, etc.

Visualizing Insights

Network Analytics

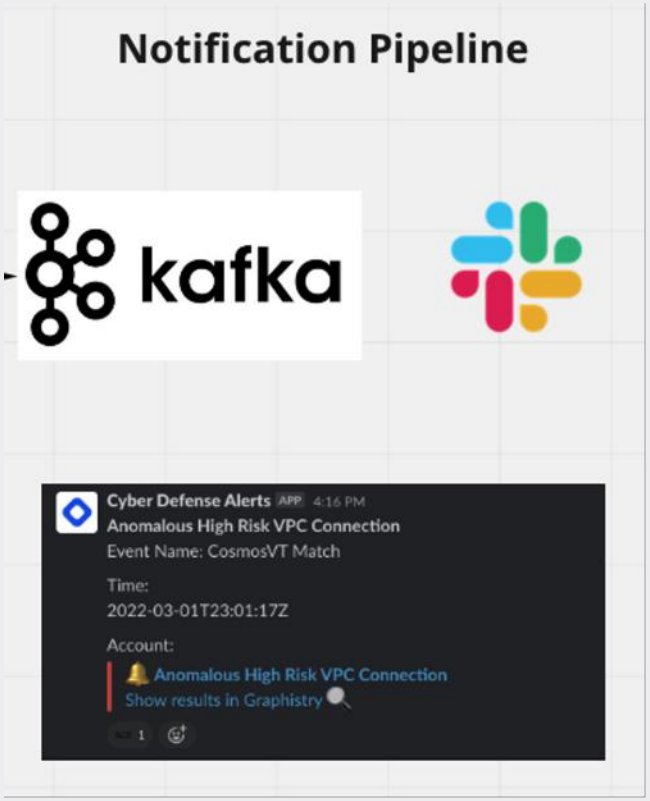
We leverage GPU-powered graph analytics to map an interactive view for our security analysts to glean insights

graphistry



Stream Alerting

Network Analytics



Our end result will be an enriched event that is streamed to a kafka topic and sent to Slack Alert to immediate notification.

The event data will also be sent to an Intel Datastore for historic storage and additional context

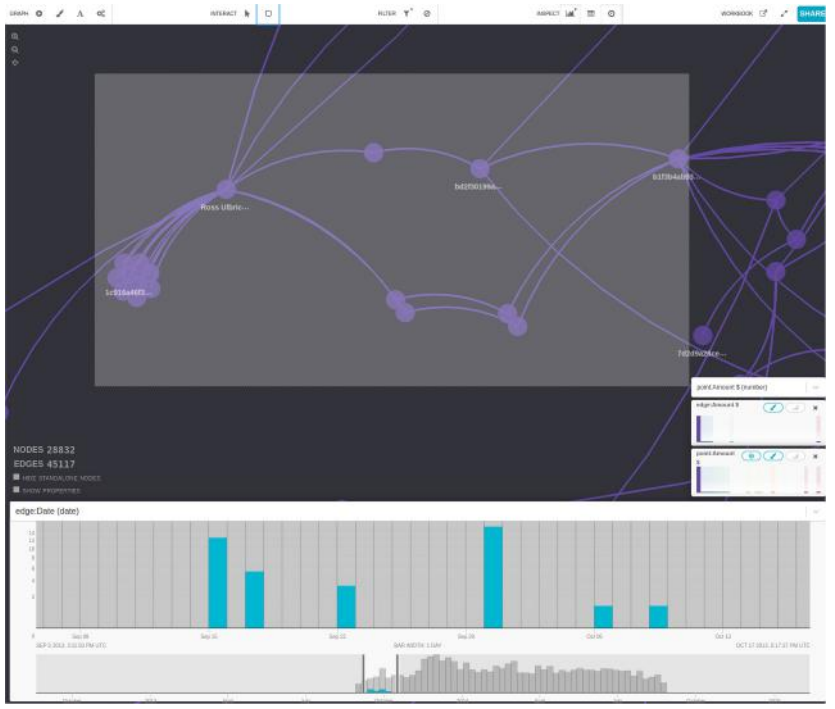
Using Graphs for Cryptotracing

Building Trust in Crypto

Ensuring compliance with OFAC

Investigation and analysis

- $1.3 billion stolen from exchanges, platforms, and private entities in the first 3 months of 2022 according to Chainalysis.
- KYC, and SAR continue to be a pain for the crypto space as wallets are generated on the fly by attackers and scammers.
- Graphs allow for companies to build out more detailed views on where funds have likely been sourced from and flag suspicious transactions or risky wallets.
- Enables for automated reporting of SAR reports and better transparency.



Graphs for Risk Quantification

Getting crypto data into Databricks

Crypto APIs

Blockchain APIs & SDKs provide a rich set of data for data science and analytics.

Services like BlockDaemon, Chainalysis, and Coinmetrics both have rich documentation with python examples.

In the case of Coinmetrics, they have a python client that returns data as a Pandas dataframe.

```
!pip install coinmetrics-api-client
```

Example Code:

```
import pandas as pd

from coinmetrics.api_client import CoinMetricsClient

api_key = '<api_key>'

client = CoinMetricsClient()

asset_metrics = client.get_asset_metrics(

    assets='btc',

    metrics=['PriceUSD', 'SplyCur'],

    start_time='2021-09-19T00:00:00Z',

    limit_per_asset=10

)

asset_metrics_df = asset_metrics.to_dataframe()
```

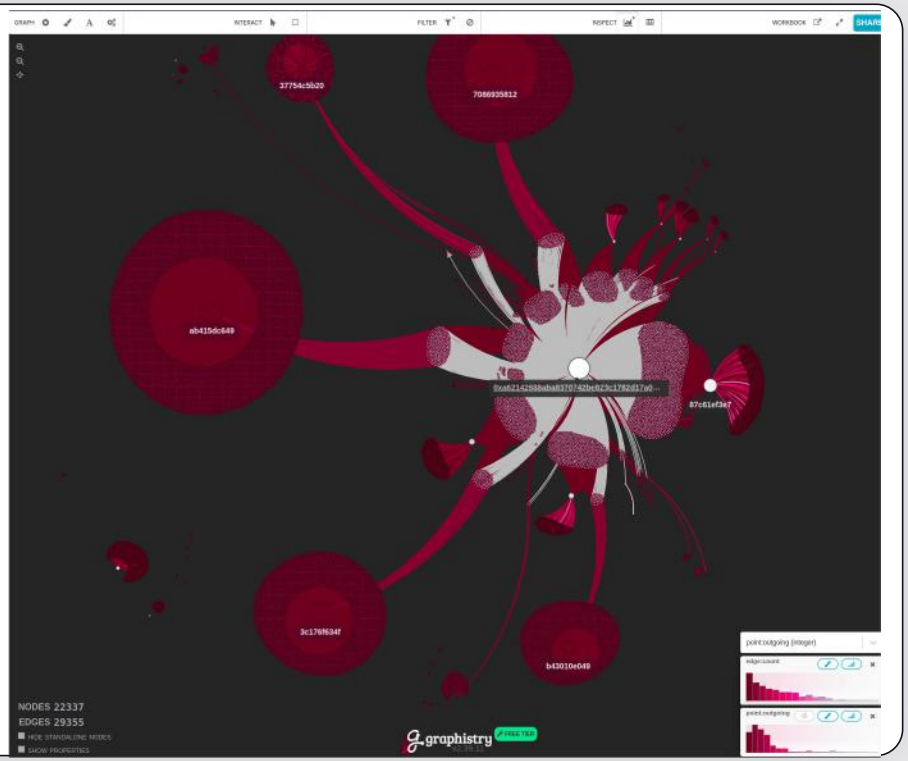
Graphs and Cryptotracing

Building out intelligence from transaction flows

Feature Selection

Services like Chainalysis, AnChain and Ciphertrace provide threat intelligence capabilities like indirect exposures to the US Treasury OFAC sanctions list.

These labels can be leveraged for a “bad neighbors” approach used in security data science (*Data Driven Security, 2014*).



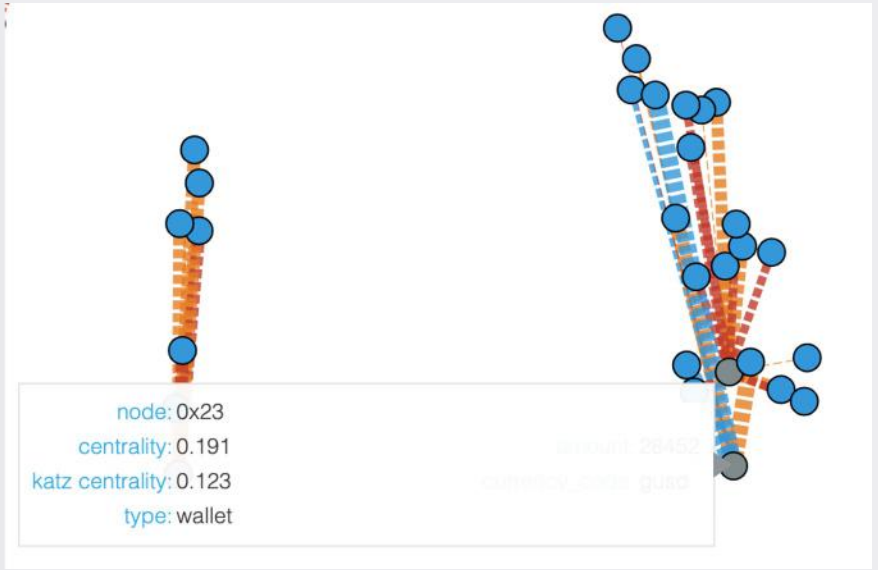
Graph Properties & Features

Centrality Measurements

Finding common properties, shared connections between wallets, exchanges or sanctioned entities can be used to quantify the risk of a completely unknown wallet.

Centrality measurements should be used to identify importance of a node's influence on the graph.

Shared wallets attackers are leveraging for referral scams, dust attacks, money laundering, and account takeover generally score higher.



Key Takeaways

Impacting business with analytics!

Focus

Find key business partners and use cases that are core to your business.

Build

Collaborate with engineering, operations, and data science teams to prototype solutions.

Adapt

Adjust architecture, models, and data strategies as needed to meet short term business objectives of your use cases.

Operationalize

Scale up on the architectures that work, measure KPIs. Communicate wins back to stakeholders to demonstrate value.

Disclaimer

Nothing in this presentation constitutes an offer to buy or sell or a solicitation of an offer to buy or sell investments, loans, securities, futures contracts, partnership interests, commodities or any other financial instruments; this presentation does not constitute, and may not be used for or in connection with, an offer or solicitation by anyone in any state or jurisdiction in which such an offer or solicitation is not authorized or permitted, or to any person to whom it is unlawful to make such offer or solicitation.

BlockFi makes no representation or warranty, express or implied, to the extent not prohibited by applicable law, regarding the advisability of investing in securities, funds, partnership interests or other investments or funding or purchasing loans. The past performance of any investment, loan, security, partnership interest, commodity or financial instrument is not a guide to future performance.

BlockFi refers to BlockFi Inc., BlockFi Lending LLC, BlockFi Trading LLC or their affiliates. BlockFi Digital Markets is a division within BlockFi and not a separate legal entity. BlockFi is not, and unless legally required, does not intend to, register as a swap dealer or futures commission merchant pursuant to the Commodities Exchange Act and the Rules and Regulations thereunder or register as a broker or dealer pursuant to the Securities Exchange Act and the Rules and Regulations thereunder.

Without limiting anything in this disclaimer, BlockFi makes no warranties and bears no liability with respect to any fund, any investments, securities, partnership interests, loans or the performance thereof.

This presentation and the views expressed in this presentation do not necessarily reflect the views of BlockFi as a whole, its directors, officers, employees, shareholders or any part or member thereof or of any third party. Nothing in this presentation constitutes, or should be construed as, investment, tax, legal, financial or any other advice.

DATA+AI SUMMIT 2022

Thank you

Anthony Tellez

BlockFi